Impact of Emotions on Fundamental Speech Signal Frequency

PAVOL PARTILA[1], MIROSLAV VOZNAK[1], ADRIAN KOVAC[2] and MICHAL HALAS[2]

[1]Department of Telecommunications
VSB-Technical University of Ostrava
17. listopadu 15/2172, 708 33 Ostrava Poruba
Czech Republic
pavol.partila@vsb.cz, miroslav.voznak@vsb.cz

[2]Department of Telecommunications
Slovak University of Technology
Ilkovicova 3, Bratislava
Slovak Republic
kovaca@ktl.elf.stuba.sk, michal.halas@stuba.sk

Abstract: - The paper deals with recognition of speeches made in a particular emotional state and examines the impact of person's emotional state on the fundamental speech signal frequency. Vocal chords create audio signals which carry information coded with human language. This process is called human speech. Based on a speech signal several speaker's attributes such as sex, age, speech disorders (stuttering or cluttering) and emotional state can be determined. As for emotions, only about 10% of speaker's emotional state or state of mind is expressed by means of speech. On that ground, a selection and a computation of suitable parameters is an important part of a system designed to determine emotions from speech signals. These parameters should be as relevant as possible in relation to the speaker's emotional state. The fundamental signal frequency is one of the speech parameters. We dealt with a method for extracting the fundamental speech signal frequency by means of a central clipping and its exploitation for the system classifying the speaker's emotional state.

Key-words: - Fundamental frequency, Emotions, Pitch extraction, Energy, Zerro Crossing Ratio, ANOVA.

1 Introduction

Man-machine interaction is a desirable trend, accompanied hand in hand with an effort to improve the quality of mutual communication. Gradually, we have achieved direct human speech interactions. On the other hand, we feel the absence of credibility of information presented by a synthetic speech from a computer's loudspeaker. Speeches generated by Text-to-Speech tools act artificially because they do not take into account the emotional state. In speech, the emotional state is characterized by specific phonetic features. These features include intensity, intonation and timbre of speech. In the domain of speech processing, the speech signals are described by parameters such as signal energy, zero crossing ratio and fundamental speech frequency or by cepstral coefficients. [2], [3] We verified the significance of the fundamental speech frequency for the determination of speaker's emotional state by applying the variance analysis [1].

2 Pre-processing

Once human speech is digitalized, the digital audio record can be analysed. In order to extract signatures such as the fundamental speech signal frequency, energy, etc., it is necessary to carry out several operations depicted in Fig. 1. These steps need to be carried out before the above-mentioned signatures have been extracted [4].

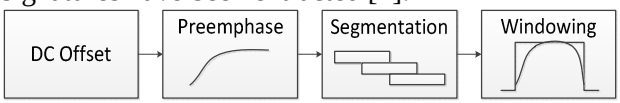

Fig. 1. Pre-processing of speech signals.

2.1 DC Offset

A number of audio cards add DC (Direct Current) components into the audio signal, as depicted in Fig. 2. Approaches used in digital signal processing are applied to compute some signatures. The DC component in the signal negatively affects the computation and may cause disturbance.

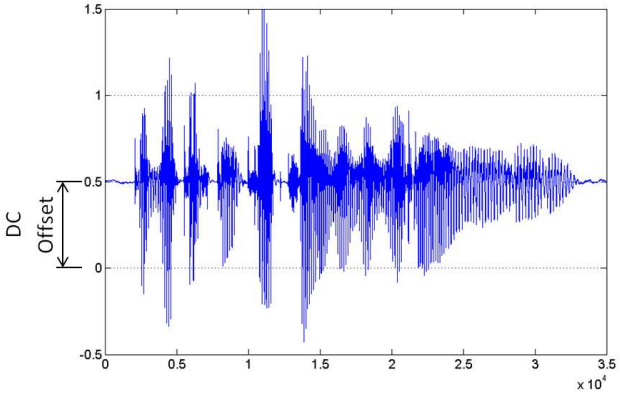

Fig. 2 Effect of DC Offset on speech signal.

It is therefore necessary to remove the DC component before the processing. The DC component of the entire signal is computed as a mean value of all analysed samples as is expressed in equation (1).

$$\mu_S = \frac{1}{N}\sum_{n=1}^{N} s(n). \quad (1)$$

The DC component is removed by a simple subtraction of the mean value.

$$s'(n) = s(n) - \mu_S. \quad (2)$$

If we do not dispose of the entire signal, typically in real-time processing when a particular part of the signal is analysed, we are not able to estimate the mean value. In this case, an online estimation of the mean value for each audio sample is used. The mean value for the current sample $\mu_s(n)$ from formula (3) can be determined once we know the mean value of the previous sample $\mu_s(n-1)$, which is linked to the actual sample by constant γ. Its value is approaching 1. The DC component is removed by a simple substraction of the mean value.

$$\mu_s(n) = \gamma\mu_s(n-1) + (1-\gamma)s(n). \quad (3)$$

2.2 Pre-emphasis

Pre-emphasis needs to be carried out due to the speech signal energy equilibration with respect to the frequency because the signal energy declines with increasing frequency. Most of speech signal energy is located in first 300Hz of the speech spectrum. Since the same important information is also included in higher parts of frequency spectrum, the so-called pre-emphasis is in most cases carried out by the FIR filter of the first order the transfer function H(z) of which is described in formula (4). The situation is depicted in Fig. 3.

$$H(z) = 1 - kz^{-1}. \quad (4)$$

Substituting the real values (5), we obtained the speech signal samples with the enhanced energy at higher frequencies. Constant k is approaching 1, i.e. it ranges from 0.95 to 1.

$$s''(n) = s'(n) - k \times s'(n-1). \quad (5)$$

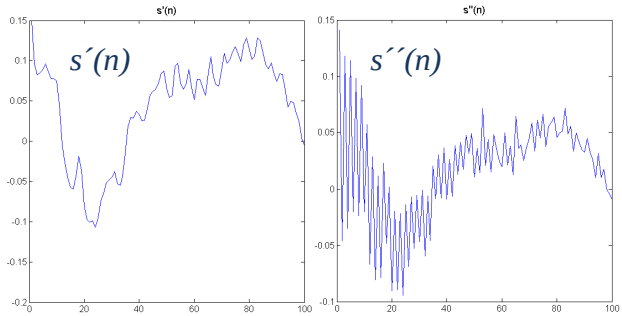

Fig. 3 Segment of speech signal before and after FIR filter preemphase.

2.3 Signal Segmentation

A speech signal is non-stationary which is not desirable from the point of view of methods for extracting parameters. Therefore, it is necessary to divide this signal into shorter segments referred to as frames. The length of these frames is chosen on the basis of the vocal tract lag. It must be short enough so that we can assume that the signal is stationary within a particular frame, but long enough at the same time. A very short frame does not have the periodic features needed for the detection of fundamental frequency ("F_0"). The frame length almost always ranges between 20 to 30 ms, giving 320 to 480 samples due to sampling frequency Fs = 16kHz. In view of the fact that the parameters of the signal in the neighbouring frames can vary rapidly it is suitable that the frames overlap, as shown in Fig. 4.

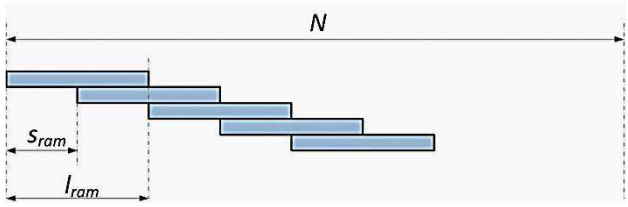

Fig. 4 Signal segmentation to frames with shifting.

The overlapping frames should not be too large or too small. In short, almost no overlap reduces the hardware difficulty, but the parameter values in neighbouring frames are very different. On the other hand, a long overlap gives smoothed waveform parameters but also increases demand for computing power and breachess the condition of independence. For this reason, in most cases, a compromise has to be made and the overlay is set up only on a half the length of one frame (10 to 15ms). To calculate the number of frames we use the floor function, i.e. rounding down. It is computed using the following equation. The variables from the equation (6) are

shown in Fig. 4. [5].

$$N_{frame} = 1 + \left\lfloor \frac{N - l_{frame}}{s_{frame}} \right\rfloor . \quad (6)$$

2.4 Smoothing Function

The segmentation of speech signals into frames results in a sharp transition at the edge. The sharp transition between the frames has an adverse effect especially when performing the frequency analysis. Multiplied window function eliminates these sharp transitions on each frame. In signal processing, many window weighted functions can be applied. For speech recognition and to avoid the impact of sharp transitions in the spectrum, the Hamming window function is used most frequently. Equation (7) contains the mathematical description describing the shape of the Hamming window. The number of samples is represented by N and n means a particular sample.

$$w(n) = 0{,}54 + 0{,}46 \cos\left[\left(\frac{1}{2}N - n\right)\frac{2\pi}{N}\right] \quad (7)$$

The Hamming window function is used because it has good spectral properties and increases the amplitude of the frame on its edges [6].

3 System Classifying Emotional State

Figure 1 shows most systems for classifying emotional state working within the given architecture. The final system consists of three blocks downwards, but it does not mean that the database of speech recordings does not play an important role. The recordings must be of good quality because they are used to debug algorithms which calculate the segmental parameters and for the train recognition system. The feature extraction in the given scheme is provided for by an algorithm using mathematical methods of signal processing which can calculate the signal energy, number of zero crossing and fundamental frequency. These methods are described in the next chapter.

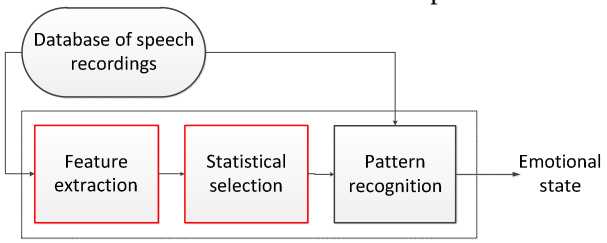

Fig. 5 System for emotional state classification.

4 Features Extraction

Once the DC component has been removed and the pre-emphasis carried out, the speech signal is ready for the extraction of the parameters. The speech recording is divided into segments (frames). Accordingly, these parameters are referred to as segmental or suprasegmental parameters and describe frame's signal features.

4.1 Speech Signal Energy

Signal energy is seen as the strength or power or voice volume. Voice energy is associated with the respiratory system. The energy of the audio signal is influenced by the conditions in which the record was created. The distance between the mouth and the microphone or lower sensitivity of hardware has a significant impact on the quality of digitalized sound. The energy parameter is calculated using the short-term signal energy which is defined in the following equation.

$$E = \frac{1}{N}\sum_{n}^{N-1}[x(n)]^2 . \quad (8)$$

The logarithm is often used to demonstrate minor changes of energy. With energy, we can determine speech activity and divide voiced and unvoiced sounds. On the other hand, signal energy sometimes cannot separate voiced and unvoiced sounds perfectly and we can lose low-energy voiced sounds.

4.2 Zero Crossing Rate

Zero Crossing Rate (ZCR hereafter) is a parameter that determines the number of zero-crossing signal levels, or in other words how many times the polarity changes. Its shape enables us to estimate the change in F_0. An increase in ZCR is associated with increasing F_0 and vice versa. ZCR carries information that is used to separate the voiced (vowels) and unvoiced (consonants, silent) sounds. The *sign* function, described in the equation below, is used to calculate ZCR.

$$ZCR(m) = \sum \left| sign[s(n)] - sign[s(n-1)] \right| \quad (9)$$

Where the number of crosses (polarity changes) is zero, the expression |sign(s(n))-sign(s(n-1))|=2.

4.3 Fundamental Frequency

As mentioned, F_0 is a very important parameter in determining the number of speech characters. F_0 carries information about the speaker, such as gender, age, speech defect or emotional state. There are several methods for detecting F_0. According to the method of calculation, they can be divided into: F_0 detection in time domain, F_0 detection in frequency domain and F_0 detection from cepstral coefficients.

ISBN: 978-1-61804-134-0

In most methods, the one-sided autocorrelation function is used, allowing us to determine the position of the first peak (from the eng. pitch extraction). The final calculation of the fundamental frequency is carried out using a simple formula (10).

$$F_0 = \frac{F_s}{k}. \quad (10)$$

To avoid detecting the fundamental frequency from the unvoiced parts of the signal it is good to set a threshold on the signal (Fig. 6 and 7). The threshold determines whether the signal is in a particular area should be voiced or unvoiced. The speech signal is very non-stationary. For this reason, it is not appropriate to set the same threshold for the whole signal. The central clipping method calculates the threshold for each frame separately.

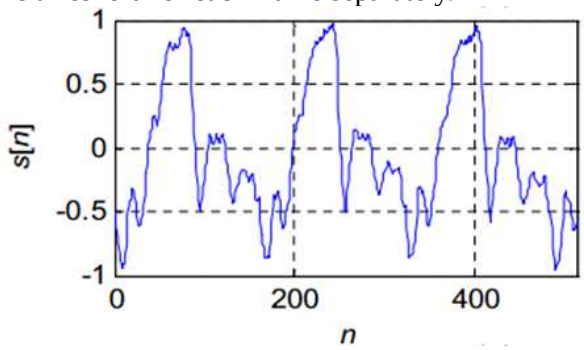

Fig. 6 One frame of speech signal before thresholding.

It determines the threshold by using the maxima from the previous and next frame, as shown in the following equation (11).

$$P_{threshold}(i) = \alpha \min(\max_{i-1}, \max_{i+1}). \quad (11)$$

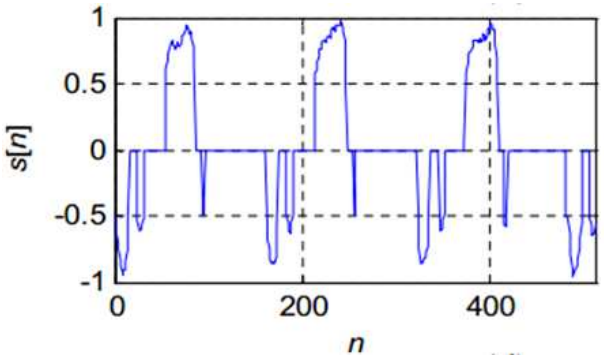

Fig. 7 One frame of speech signal after thresholding.

Where Pi is the threshold for frame i. Parameter α is approximately 0.8. The size of the sample of the signal is then weighted with this threshold and normalized to values of 1, 0 and -1 as is depicted in Fig. 8.

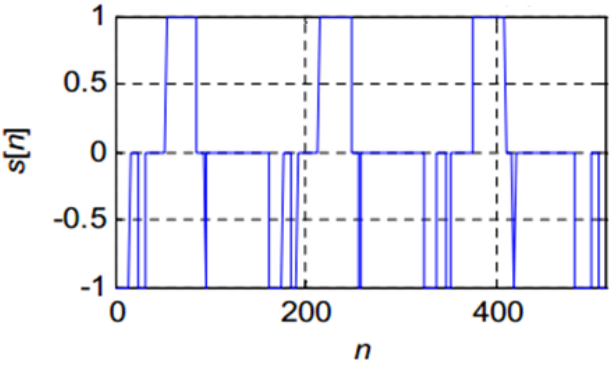

Fig. 8 Normalized frame with amplitudes 1, 0, -1.

The following autocorrelation of such normalized signal reveals the position of the peak. Using this position, F_0 is calculated in relation (8).

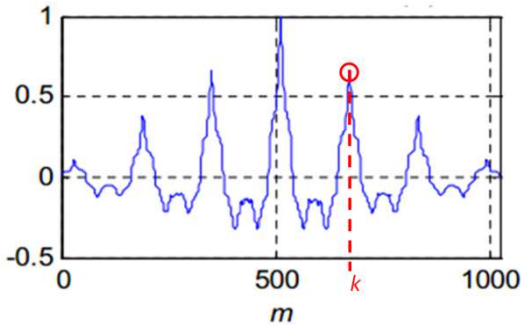

Fig. 9 Pitch extraction from autocorrelated voiced frame of speech.

The calculation of the fundamental frequency of the speaker is dependent on the method used. In some cases, the chosen method may not be the most accurate [7].

5 Statistical Selection

Before the algorithm for the final classification of speaker's emotional state is developed and trained, it is necessary to verify the importance of a particular parameter. A parameter that is not significant for researched information – in this case emotional state – should not be used to train a classifier. The verification of the significance of the parameter gives us the opportunity to make a conclusion on the correctness of the computation method applied.

5.1 Data Analysis

The theory of language processing says that the energy and zero crossings rate are important parameters. Where the methods of calculation are correct, it is not necessary to establish their significance. On the other hand, the method for calculating F_0 may be dependent on the characteristics of speech recordings. For this reason,

ISBN: 978-1-61804-134-0

it is appropriate to check whether the extracted F_0 parameter is statistically significant.

The algorithm for F_0 extraction was designed in Matlab version R2010b. Data analysed is F_0, calculated from recordings of different emotional states. These recordings were recorded by 5 women. This research was narrowed to three types of emotional state: anger, fear and neutral emotional state. These three emotional states were selected on the assumption that the parameter F_0 for them is more significant than for other emotional states. The study population consists of three sets: 1_neutral (30 records of 5 speakers in a neutral state), 2_anger (30 records of 5 angry speakers) and 3_fear (30 records of 5 fearful speakers).

6 Analysis of Variance

The analysis of variance (ANOVA) allows some comparisons of mean values of independent choice. As F_0 is the data set for different speech recordings, we consider them a set of independent choice. ANOVA in its parametric form implies two conditions: Normality , the distributions of the residuals are normal and Homoscedasticity, the variance of data in groups should be the same. For this reason it is necessary to verify these conditions [8], [9].

6.1 Tests for Homoscedasticity

The H_0 hypothesis holds: The variances of the three groups (1_neutral, 2_anger, 3_fear) selection are identical. This hypothesis is verified by three tests.

	Test	P - value
Cochran's	0,54007	0,00815
Bartlett's	1,10384	0,01385
Levene's	4,31538	0,01630

Table 1 Cochran's, Bartlett's and Levene's tests of homoscedasticity for groups 1_normal vs. 2_anger vs. 3_fear.

The H_0 hypothesis is rejected because P-value of the tests from Tab. 1 is less than 0.05. This means that the variances of the individual selections are not identical. Therefore, it is only possible to compare the following F_0: anger versus the neutral state; and fear versus the neutral state.

	Test	P-value
Cochran's	0,66698	0,05550
Bartlett's	1,05560	0,06710
Levene's	2,54559	0,11561

Table 2 Cochran's, Bartlett's and Levene's tests of homoscedasticity groups 1_normal vs. 2_anger.

	Test	P-value
Cochran's	0,58629	0,38931
Bartlett's	1,01523	0,38931
Levene's	1,82875	0,18236

Table 3 Cochran's, Bartlett's and Levene's tests of homoscedasticity groups 1_normal vs. 3_fear.

P-values of tests from Tab. 2 and 3 are greater than 0.05 so we can accept the H_0 hypothesis with a 95% probability. This means that F_0 recordings 1_neutral 2_anger and have identical variances. The same result applies to recordings of F_0 1_neutral versus 3_fear [8].

6.2 Normality Test

The normality test should be performed for each researched selection (for each emotional state). Where the size of P-value of the Kolmogorov-Smirnov test exceeds 0.05, we can conclude that the range studied may be subject to the normal distribution.

Kolmogorov–Smirnov test	P-value
1_neutral	0,904292
2_anger	0,241772
3_fear	0,860042

Table 4 Test of normality for each selection (emotional state).

P-value for the three groups is more than 0.05 which means that all three selections have a normal distribution [8].

6.3 ANOVA

Once the conditions of homoscedasticity and normality have been verified, it is possible to perform analysis of variance. Test statistics applied within the analysis of variance is F-ratio, which was derived from the basic analysis of variability of input data files. F-ratio is calculated as the ratio of the mean square between groups MSB and the mean square within groups MSW as described in equation (12). The calculation of the mean squares is not described. It can be found in references [8], [9].

$$F_{ratio} = \frac{MS_B}{MS_W}. \tag{12}$$

$$P_{value} = 1 - F_{ratio}. \tag{13}$$

F-ratio statistics is sensitive to the validity of the H_0 hypothesis which is defined as equality of mean values of the selections studied. H_0: Mean values of F_0 for the selections are equal.

ISBN: 978-1-61804-134-0

$$\mu_{F_0}^{A} = \mu_{F_0}^{B}. \quad (14)$$

ANOVA_1_2	Mean square	F-ratio	P-value
Between gruops	7647,07	23,92	0,00
Within groups	319,66		

Table 5 Analysis of variance for 1_neutral and 2_anger.

ANOVA_1_3	Mean square	F-ratio	P-value
Between gruops	26011,80	153,13	0,00
Within groups	169,86		

Table 6 Analysis of variance for 1_neutral and 3_fear.

P-value is less than the specified level of significance (0.05). Thus, the H_0 hypothesis was rejected. Mean values of the F_0 of the vocal tract in these emotional states are not the same. The method used to extract the basic tone is correct to a certain extent. This argument can also be estimated from the Fig. 10.

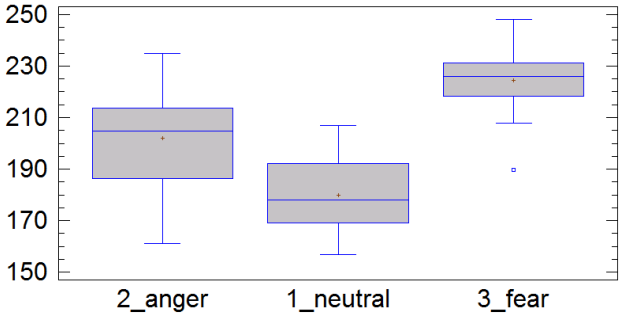

Fig. 10 Fundamental frequencies for F_0 extracted from recordings of 3 emotional states.

7 Conclusion

Choosing the method for calculating the fundamental frequency of the human voice is a very important step which should precede the classifier's designing and training. It was therefore necessary to check the statistical significance of this parameter. The average tone of all speakers was established as 202.12 Hz. The maximum value obtained was for the emotional state of fear (248Hz), the minimum in the neutral state (156Hz). The exploratory statistics and box plot for all three emotional states show that the speakers' F_0 increased most in the state of fear. The fundamental tone of speakers in the neutral state was low. The ANOVA test established that the condition for the homoscedasticity comparison is not fulfilled for all three selections because the variances are not the same. Therefore, it was better to compare the two emotional states to the neutral state.

The ANOVA test confirmed that the mean F_0 emotional state of anger and fear are not equalled as F_0 for the neutral state, thus confirming the assumption that a change in the speaker's emotional state changes the fundamental frequency of your vocal tract.

The conclusion is that if we had this single parameter, the final classifier would only be able to detect speeches other than those made in a neutral state of mind. To be able to classify a particular emotional state (fear, anger, happiness, sadness, ...), it is necessary to obtain more parameters to be enable the processing of human language and follow up on their mutual correlations.

Acknowledgement

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 218086 and from project "Prediction, modelling and assurance of quality of VoIP voice services" supported by the Slovak University of Technology "Grant programme to support young researchers".

References:

[1] J. Psutka, J. Muller and V. Radova, *Mluvíme s počítačem česky*. Praha: Academia, 2006.

[2] M. Schroder, *Emotional Speech Synthesis: A Review*. Proc. Eurospeech 2001, ISCA, Bonn, Germany, 2001,

[3] D. Ververidis, D. Kotropoulos and I. Pitas, *Automatic Emotional Speech Classification.* Quebec: ICASSP, 2004. pp. 593-596.

[4] D. Gerhard, *Pitch extraction and fundamental frequency*: *history and current techniques*. University of Regina, 2003.

[5] J. Sole-Casals, and V. Zaiats, Advances in nonlinear speech processing, In *International Conference on Nonlinear Speech Processing, NOLISP 2009*, New York: Springer, 2010.

[6] J. Picone, *Signal Modeling Techniques in Speech Recognition.* In Proceeding of the IEEE, 1993.

[7] K. Kavita and S. Zahorian, *Yet Another Algorithm for pitch Extraction*. Florida: ICASSP, 2002. pp. 361-364.

[8] M. Roberts and J. R. Russo, A student's guide to analysis of variance, In *Proc. of the 7th International Conference on Autonomous Agents and multiagent systems*, Estoril, Portugal, 2008.

[9] R. Bris, M. Litschmannova, *Statistika II*, College book, VSB–Technical University of Ostrava, 2007.